A Indeterminação da Autoria: A ModelagemSociodemográfica e os Limites da Inteligência Artificialna Previsão da Redação do ENEM no Estado do Rio deJaneiro

Francisco Alves de Freitas Neto, Carlos Henrique Medeiro de Souza, Fabio Machado de Oliveira, Leonard Barreto Moreira

https://doi.org/10.1590/SciELOPreprints.17449

Submetido em: 2026-08-17
Postado em: 2026-08-18 (versão 1)
(AAAA-MM-DD)

A Indeterminação da Autoria: A Modelagem Sociodemográfica e os Limites da Inteligência Artificial na Previsão da Redação do ENEM no Estado do Rio de Janeiro.

The Indeterminacy of Authorship: Sociodemographic Modeling and the Limits of Artificial Intelligence in Predicting ENEM Essay Scores in the State of Rio de Janeiro.

Francisco Alves de Freitas Neto (IFF / UENF) [1], Carlos Henrique Medeiro de Souza (UENF) [2], Fabio Machado de Oliveira (UENF) [3], Leonard Barreto Moreira (UFF) [4]

Resumo - Este trabalho investiga os limites teóricos e operacionais da Inteligência Artificial na predição do desempenho discursivo na avaliação das notas de redação do ENEM, articulando a Modelagem de Sistemas de Inteligência Artificial e a Análise do Discurso. A hipótese central postula que modelos preditivos alimentados por *proxies* sociodemográficos são ineficazes para mapear completamente o letramento estrutural, principalmente nos espaços de desempenho mediano, categoria onde a agência polifônica e a autoria se sobrepõem ao determinismo material. O percurso metodológico consistiu no desenvolvimento de modelos de *Machine Learning* para classificar sujeitos em três estratos analíticos ("Abaixo", "Acima"e "Mediana"). A partir de um modelo *Baseline* estruturado em *Random Forest*, implementou-se uma rede neural profunda *Multilayer Perceptron* (MLP) otimizada. Para garantir o rigor da avaliação, a base de treinamento foi equalizada sinteticamente pelo algoritmo SMOTE-NC nas três classes. Diante da falha algorítmica inicial em identificar o estrato mediano no mundo real, foram aplicadas técnicas de Justiça Algorítmica (*Fairness AI*), inicialmente por meio da penalização assimétrica do erro via *Cost-Sensitive Learning* e, posteriormente, através do ajuste de pós-processamento com *Threshold Moving*. Os resultados empíricos demonstraram que o modelo manteve forte tração determinista nos extremos socioeconômicos, mas sofreu com falta de acurácia na classe central. As calibrações éticas resultaram apenas na redistribuição matemática do erro e no aumento da entropia, limitando a acurácia global a um teto de aproximadamente 54%. Conclui-se que o desempenho discursivo mediano possui propriedades não-lineares que resistem à redução matricial. A estagnação preditiva da máquina fornece evidências empíricas de que a autoria humana, quando submetida a condições de igualdade material, apresenta-se como de difícil predição computacional.

Palavras-chave: Inteligência Artificial; Análise do Discurso; Redes Neurais; Justiça Algorítmica; Função-Autor.

Abstract - This work investigates the theoretical and operational limits of Artificial Intelligence in predicting discursive performance in the assessment of ENEM essay scores, integrating Artificial Intelligence Systems Modeling and Discourse Analysis. The central hypothesis postulates that predictive models fed by sociodemographic *proxies* are ineffective in fully mapping structural literacy, particularly within

[1]Mestre em Pesquisa Operacional e Inteligência Computacional - UCAM - ORCID: https://orcid.org/0000-0001-8359-004X
[2]Doutor em Comunicação - UFRJ - ORCID: https://orcid.org/0000-0002-3774-0323
[3]Doutor em Cognição e Linguagem – UENF - ORCID: https://orcid.org/0000-0003-1336-2994
[4]Doutor em Modelagem Computacional - UERJ - ORCID: https://orcid.org/0000-0002-5588-2923

spaces of median performance—a category where polyphonic agency and authorship override material determinism. The methodological approach consisted of developing *Machine Learning* models to classify subjects into three analytical strata ("Below", "Above", and "Median"). Building upon a *Baseline* model structured on *Random Forest*, an optimized *Multilayer Perceptron* (MLP) deep neural network was implemented. To ensure evaluation rigor, the training dataset was synthetically equalized across the three classes using the SMOTE-NC algorithm. Given the initial algorithmic failure to identify the median stratum in real-world data, *Fairness AI* techniques were applied, initially through asymmetric error penalization via *Cost-Sensitive Learning*, and subsequently through a post-processing adjustment using *Threshold Moving*. Empirical results demonstrated that the model maintained strong deterministic traction at the socioeconomic extremes, yet suffered from a lack of accuracy in the central class. Ethical calibrations merely resulted in the mathematical redistribution of the error and an increase in entropy, limiting the overall accuracy to a ceiling of approximately 54%. It is concluded that median discursive performance possesses non-linear properties that resist matricial reduction. The machine's predictive stagnation provides empirical evidence that human authorship, when subjected to conditions of material equality, presents itself as computationally difficult to predict.

Keywords: Artificial Intelligence; Discourse Analysis; Neural Networks; Fairness AI; Author-Function.

1 Introdução

O Estado do Rio de Janeiro configura-se como um território atravessado por divisões socioespaciais e culturais importantes, nas quais a distribuição de capital econômico e infraestrutura obedece a uma lógica de desigualdade histórica. Observa-se um contraste entre os polos de riqueza da Região Metropolitana e os municípios fluminenses mais afastados, como aqueles localizados na Baixada e no extremo Noroeste Fluminense. Essa heterogeneidade não é apenas geográfica, mas também determina o acesso a bens e serviços fundamentais, constituindo o substrato socioeconômico sobre o qual as trajetórias de desenvolvimento individual e cognitivo são traçadas.

No âmbito educacional, essa assimetria social reflete-se diretamente no cenário do Ensino Médio. Estudantes da rede pública, especialmente aqueles inseridos em contextos periféricos ou no interior do estado, enfrentam não apenas a precarização estrutural das instituições, mas também o acesso restrito aos equipamentos culturais e às elites de letramento hegemônicas. O sistema escolar, em vez de atenuar essas disparidades, frequentemente as consolida, fazendo com que o percurso formativo seja mediado pelas condições materiais de existência e limitando a apropriação do capital cultural validado pelas instituições de prestígio. Essa dinâmica evidencia o que Frigotto et al. (2012) aponta como a dualidade educacional intrínseca ao modo de produção, na qual a educação permanece dividida entre a precarização destinada às classes trabalhadoras e a apropriação do capital cultural reservada aos grupos dirigentes.

É nesse contexto de exclusão estratificada que o Exame Nacional do Ensino Médio (ENEM) se estabelece como a principal via de acesso ao ensino superior. Contraditoriamente concebido como um instrumento de democratização, o exame, incluindo sua prova de redação, atua como um dispositivo de filtragem discursiva. A avaliação do texto dissertativo-argumentativo exige a mobilização de um repertório sociocultural específico e o domínio estrito da norma-padrão. Para o estudante fluminense afastado dos centros de poder sociolinguístico, a nota da redação vai além da mera avaliação de uma aptidão cognitiva abstrata; ela quantifica sua distância geográfica e o acesso à recursos educacionais em relação ao centro linguístico da capital.

Com a crescente adoção de tecnologias de inteligência artificial na análise de políticas educacionais, há uma necessidade cada vez maior de aplicar algoritmos de Aprendizado de Máquinas (*Machine Learning*) para modelar e prever o desempenho acadêmico (Alyahyan; Düştegör, 2020; Rastrollo-Guerrero et al., 2020). Ao alimentar esses modelos com microdados socioeconômicos e territoriais para antecipar variáveis complexas, como a nota da redação, adentra-se em um território incerto. A máquina,

regida por uma lógica de otimização probabilística, não analisa a enunciação ou a subjetividade do candidato, mas identifica correlações estatísticas entre vulnerabilidade material e notas que historicamente são mais baixas. O deficit de políticas públicas é assim codificado pelo algoritmo, convertendo-se em uma métrica de limitação individual que restringe a ”função-autor” a um determinismo matemático.

Diante desse cenário de transição tecnológica e da potencial automação de dinâmicas excludentes, a presente pesquisa se estrutura a partir da seguinte questão central: De que maneira o atravessamento da Inteligência Artificial na predição de notas do ENEM tensiona a redução da subjetividade e da autoria do candidato fluminense a meras correlações estatísticas, especialmente na zona de indeterminação social da classe de notas medianas?

Para nortear este trabalho, a pesquisa estabelece como objetivo central investigar o processo pelo qual algoritmos de aprendizado de máquina, especificamente as Redes Neurais Artificiais (RNAs), adquirem, automatizam e prenunciam incertezas ao modelar uma classificação da nota da redação do ENEM. O foco recai sobre a análise de como as profundas disparidades socio-territoriais do Rio de Janeiro, materializadas nos microdados do INEP na forma de variáveis qualitativas socioculturais e territoriais, são convertidas em incertezas preditivas. Dessa forma, busca-se compreender os limites da máquina ao reduzir a complexidade da cognição humana e do letramento a correlações estatísticas que frequentemente espelham desigualdades estruturais.

Para alcançar essa meta, o percurso metodológico desdobra-se em três frentes específicas. Inicialmente, propõe-se a análise do perfil estatístico descritivo da base de dados referente ao Estado do Rio de Janeiro, mapeando como as variáveis sociodemográficas estabelecem e denunciam, historicamente, tendências de exclusão que precedem qualquer treinamento computacional. Em seguida, o estudo busca comparar o grau de absorção e reprodução desses vieses socioculturais entre diferentes arquiteturas preditivas, contrapondo o desempenho das RNAs ao de modelos baseados em grupos de árvores de decisão, especificamente o algoritmo *Random Forest*. Por fim, a pesquisa visa avaliar a eficácia dos mecanismos técnicos de mitigação de viés algorítmico (*fairness algorithms*) aplicados a esse cenário, problematizando de forma crítica se as correções matemáticas inseridas na arquitetura computacional são suficientes para mitigar erros preditivos enraizados nas dinâmicas sociais do estado.

2 Fundamentação Teórica

A predição do desempenho acadêmico por meio de ferramentas computacionais tornou-se um dos campos mais importantes da mineração de dados educacionais. Conforme mapeado por Rastrollo-Guerrero et al. (2020) em sua revisão sistemática, a última década presenciou uma proliferação de modelos de *Machine Learning* projetados para antecipar trajetórias estudantis e subsidiar intervenções pedagógicas. Contudo, ao transpor esse cálculo algorítmico para o contexto brasileiro, especificamente para o Exame Nacional do Ensino Médio (ENEM), a predição deixa de ser um mero exercício estatístico de acertos em provas e passa a operar como um reflexo das desigualdades estruturais do país. Como argumentam Bodart e Coimbra (2020), o ENEM atua historicamente não apenas como um instrumento de avaliação, mas também como um mecanismo de reprodução de desigualdades sociais, no qual o capital cultural e econômico prévio define a fronteira do sucesso acadêmico.

Essa reprodução é mensurável e tem sido documentada na literatura por meio da análise dos microdados do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP). Estudos como o de Pires (2017) demonstram que variáveis como a renda familiar e a escolaridade dos pais exercem um peso determinante nas notas, estabelecendo uma barreira para candidatos de classes populares. Esse determinismo socioeconômico é agravado pela cisão institucional do ensino básico nacional, onde a diferença de desempenho entre jovens de escolas públicas e privadas fundamenta a

exclusão educacional (Feijó; França, 2020). Mais recentemente, Xavier e Rodrigues (2024) adicionaram uma camada geo-tecnológica a essa equação, revelando que a desigualdade regional e a lacuna de conectividade digital (o acesso à infraestrutura de internet) atuam como tensores que aprofundam as assimetrias de longo prazo nos resultados do ENEM.

2.1 A Complexidade de se Avaliar e Quantificar o Valor de uma Redação

Avaliar um texto é, sem dúvida, uma tarefa complexa. Quando se observa as suas enunciações, o texto oferece os indícios necessários para a compreensão do sujeito que o escreve. Bardin (2011), ao estruturar as bases da Análise de Conteúdo, demonstra que o processo analítico parte da mensagem manifesta para, através de inferências, compreender as condições de produção sociológicas, históricas e psicológicas do enunciador. Trata-se de um movimento hermenêutico e indutivo que caminha do texto para o contexto, permitindo que o leitor compreenda a visão de mundo do aluno a partir de sua comunidade e inserção sociocultural.

O que a modelagem algorítmica preditiva propõe, contudo, é a inversão desse caminho epistemológico. Ao treinar um algoritmo de *Machine Learning* (ML) com variáveis sociodemográficas para prever a nota de uma redação, assume-se a premissa de que o percurso inverso também seria igualmente válido. Se, por um lado, o texto reflete a realidade material de quem o escreve, por outro, é metodologicamente inviável deduzir a qualidade cognitiva, a complexidade argumentativa e a criatividade discursiva de um autor exclusivamente a partir de marcadores estatísticos de classe ou território. A enunciação é um evento de sistema aberto; portanto, o valor intrínseco de uma redação não pode ser decodificado a priori pelas coordenadas de um questionário socioeconômico.

Ao buscar quantificar o valor de uma redação, os avaliadores tentam enquadrar a competência de um sujeito empírico e mensurar uma suposta capacidade discursiva. Contudo, essa tentativa esbarra na própria ontologia da autoria. Como Barthes (2004) postulou, a escritura exige a ”morte do autor”, sendo o espaço onde a voz do indivíduo empírico se dilui em favor da própria linguagem. A avaliação automatizada ignora essa diluição ao tentar vincular o texto aos marcadores demográficos do sujeito.

Foucault (1997)) expande essa crítica ao argumentar que o autor não deve ser visto como um indivíduo isolado, criador de uma obra, mas sim como uma ”função discursiva” (a função-autor) que opera em sistemas específicos de produção e regulação. Essa função é responsável por definir as formas de recepção, interpretação e legitimação de um texto, conferindo-lhe validade com base em critérios historicamente estabelecidos. No contexto do ENEM, o algoritmo atua como um regulador impreciso dessa função, negando a legitimidade autoral a sujeitos que não se enquadram no perfil demográfico estatisticamente associado ao letramento hegemônico.

Adicionalmente, a predição algorítmica ignora a natureza dialógica da linguagem. Fiorin (2016) apoia a ideia da dissolução da figura do autor, não apenas por suas funcionalidades institucionais, mas também por uma essência estrutural particular, exclusiva de sua vivência, ao introduzir os conceitos de polifonia de Bakhtin no cenário enunciativo. Para ele, o autor é compreendido como um conjunto de vozes sociais que competem pelas intenções do texto. Na continuidade dessa tradição, Authier-Revuz (1990) aprofunda a noção de heterogeneidade enunciativa, destacando a tensão intrínseca entre a pretensão do sujeito de se constituir como fonte autônoma do sentido e a realidade de que ”toda a fala é determinada de fora da vontade do sujeito e que este é mais falado do que fala” (Authier-Revuz, 1990, p. 26).

Essa perspectiva evidencia que, mesmo quando parametrizada pelo tema escolhido e pelo estruturalismo discursivo de um texto dissertativo e argumentativo, o autor candidato fluminense é influenciado por outras vozes e por suas condições socioculturais. o ML, no entanto, desprovido da capacidade de lidar com essa complexidade polifônica e a heterogeneidade constitutiva, converte os microdados

em tensores matemáticos. A máquina não avalia a qualidade do engajamento discursivo ou a multiplicidade de vozes orquestradas no texto; ela simplesmente achata a função-autor, transformando as condições externas que ”falam” sobre o sujeito em uma sentença matemática de exclusão prévia.

2.2 A Arquitetura Avaliativa do INEP: As Cinco Competências da Redação e o Letramento Hegemônico

A nota da redação do Exame Nacional do Ensino Médio (ENEM) é o resultado de uma matriz de referência dividida em cinco eixos avaliativos. Conforme estruturado no Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2024), a nota final do candidato corresponde à soma do desempenho obtido nessas cinco competências específicas (vide Quadro 1). Do ponto de vista da cognição e da linguagem, essas métricas têm o objetivo de quantificar desde a adequação gramatical até a complexidade argumentativa e a capacidade de proposição social.

Quadro 1 – Competências avaliadas na redação do ENEM

Competência 1:	Demonstrar domínio da modalidade escrita formal da língua portuguesa.
Competência 2:	Compreender a proposta de redação e aplicar conceitos das várias áreas de conhecimento para desenvolver o tema, dentro dos limites estruturais do texto dissertativo-argumentativo em prosa.
Competência 3:	Selecionar, relacionar, organizar e interpretar informações, fatos, opiniões e argumentos em defesa de um ponto de vista.
Competência 4:	Demonstrar conhecimento dos mecanismos linguísticos necessários para a construção da argumentação.
Competência 5:	Elaborar proposta de intervenção para o problema abordado, respeitando os direitos humanos.

Fonte: (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024, p. 6).

Quando a pesquisa propõe o treinamento de uma Rede Neural Artificial (RNA) ou de algoritmos preditivos, como o *Random Forest*, utilizando as bases do INEP, busca-se observar diretamente as cinco competências apresentadas no Quadro 1, que funcionam como variável alvo (*target*). O risco epistemológico e metodológico reside no fato de que o algoritmo pode não ser capaz de perceber a complexidade envolvida na avaliação do texto, uma vez que as variáveis complementares dos Microdados do INEP não revelam informações sobre aspectos como a estratégia argumentativa construída na Competência 3 ou a ética exigida na Competência 5.

Portanto, o que a máquina faz é buscar padrões estatísticos entre as variáveis de entrada e as pontuações numéricas obtidas nessas cinco matrizes. Ao fazê-lo, a Inteligência Artificial observa o abandono histórico do capital cultural fluminense em favor de uma elite, mapeando como a escassez material muitas vezes reflete no distanciamento da norma-padrão (Competência 1) e do repertório enciclopédico (Competência 2), e automatizando matematicamente essa correlação como um destino provável.

Aprofundando essa divisão ontológica entre a máquina e o sujeito, Morin e Le Moigne (2000) oferecem uma visão epistemológica mais abrangente. Ao postular sobre a complexidade da inteligência, Morin adverte que a cognição humana, a linguagem e a sociedade são sistemas inerentemente abertos e hipercomplexos, caracterizados por uma altíssima interdependência entre suas partes e o ambiente (ecossistêmicos). As máquinas, por outro lado, são sistemas fechados e isolados, desprovidas de inserção existencial. A inteligência artificial não ”compreende” o texto da redação do ENEM ou o peso da geografia fluminense; ela apenas computa regularidades dentro de um universo algorítmico probabilístico, enviesado pela escolha da base, sua classificação e a própria arquitetura de codificação.

> (...) esses modelos negligenciam uma das características fundamentais da complexidade, sublinhada por H. A. Simon: a representação – e portanto a eventual medida – da complexidade de um sistema (ou de uma mensagem) é inteiramente baseada num esquema de codificação ("*the encoding scheme*"). O código muda e o modelo mudará e por conseguinte o valor da medida. Se a escolha do código pertence ao modelizador, não podemos, portanto, pretender a objetividade da medida (Morin; Le Moigne, 2000, p. 238).

Essa observação descreve o que ocorre ao tentar prever a nota da ”Competência 3” (vide Quadro 1): a capacidade de selecionar, relacionar e defender um ponto de vista a partir de variáveis. O algoritmo, restrito ao paradigma de regras fechadas, não apresenta incertezas devido a um defeito de implementação ou à falta de dados; erra porque a própria tarefa de modelar a ”função-autor” a partir de metadados sociodemográficos exige a simplificação de um sistema hipercomplexo e ecossistêmico em um micro-domínio fechado, replicando exatamente a limitação estrutural do aprendizado de máquinas.

Portanto, quando a IA tenta modelar a cognição e o letramento de um aluno a partir de seus dados socioeconômicos, comete uma simplificação epistemológica de algo mais complexo: confina a abertura dialógica e a interdependência do sujeito a uma matriz probabilística de pesos e vieses. O fracasso preditivo e a consequente automação da exclusão não são erros corrigíveis, mas sim sintomas da incapacidade de um sistema fechado em lidar com a complexidade do mundo vivo.

2.3 Variáveis Proxy e ”Profecias Autorrealizáveis: O Algoritmo como Arma de Destruição Matemática”

Se a complexidade ecosistêmica humana escapa à captação por algoritmos fechados, a solução encontrada para adaptar a realidade nos modelos estatísticos é o reducionismo. É neste ponto que a análise sociotécnica desenvolvida por O’Neil (2020) demonstra que, quando modelos preditivos são aplicados em áreas sensíveis, como educação e crédito, utilizando dados opacos e inquestionáveis, eles se tornam ”Armas de Destruição Matemática” (ADMs), ferramentas que punem os mais vulneráveis e favorecem uma elite, sob a fachada da objetividade científica.

O primeiro ponto de convergência com a limitação algorítmica é o uso massivo do que O’Neil define como variáveis *proxy* (substitutas) e o impacto do viés de representação nos bancos de treinamento. Como a Inteligência Artificial é incapaz de enxergar a complexidade cognitiva, a polifonia ou o verdadeiro letramento de um candidato no texto da redação, busca correlações matemáticas em dados periféricos. No cenário fluminense analisado, os microdados revelam que o viés social atua por meio da assimetria de amostragem: os extremos sociais tendem a ser mais afetados do que a maioria dos candidatos que frequentam os estratos centrais. Nestes extremos sociais, os preditores matemáticos que efetivamente ancoram os vieses recaem sobre marcadores estruturais, tais como a renda familiar, a cor/raça e a rede de ensino, com ampla vantagem de desempenho atribuída às escolas particulares e a cor branca da pele. A máquina, alheia à sociologia, converte a materialidade dessa exclusão em matrizes de probabilidade preditiva. Dessa forma, a RNA reforça a desigualdade socioeconômica e

racial não como um deficit de políticas públicas, mas como uma pretensa incapacidade redacional e cognitiva do sujeito.

A gravidade dessa substituição de variáveis culmina no segundo conceito central de (O'Neil, 2020): a criação de ciclos de *feedback* destrutivos, também conhecidos como profecias autorrealizáveis. Ao contrário de modelos científicos robustos, que são ajustados quando cometem erros, as ADMs educacionais constroem a realidade que elas mesmas preveem. Se o algoritmo, treinado com base no histórico do ENEM, prevê notas mais baixas para alunos de escolas públicas de baixa renda familiar, esses resultados preditivos podem direcionar políticas que os consideram "investimentos de alto risco". Ao receberem menos incentivos, nivelamentos ou acesso a tecnologias de tutoria, esses mesmos alunos inevitavelmente apresentarão um desempenho inferior no exame oficial. Essa nova nota baixa retroalimentará a base de dados do INEP no ano seguinte, "provando" matematicamente à IA que a sua previsão excludente estava absolutamente correta. A exclusão deixa de ser um fenômeno sociológico para se tornar um teorema validado pela própria máquina.

3 Metodologia

Para investigar como os algoritmos reconhecem os padrões das desigualdades estruturais codificadas nos microdados do INEP, o percurso metodológico desta pesquisa foi estruturado em três etapas analíticas e experimentais complementares. O *dataset* empírico é constituído pelos microdados do Exame Nacional do Ensino Médio (ENEM) referentes ao ano de 2024, extraídos do repositório oficial da plataforma do Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (INEP)[5]. Portanto, o desenho da pesquisa parte da premissa de que a avaliação do letramento realizada pelo algoritmo ocorre em uma análise social, racial e geográfica profunda, exigindo uma abordagem que tensione a estatística em relação à realidade material dos candidatos.

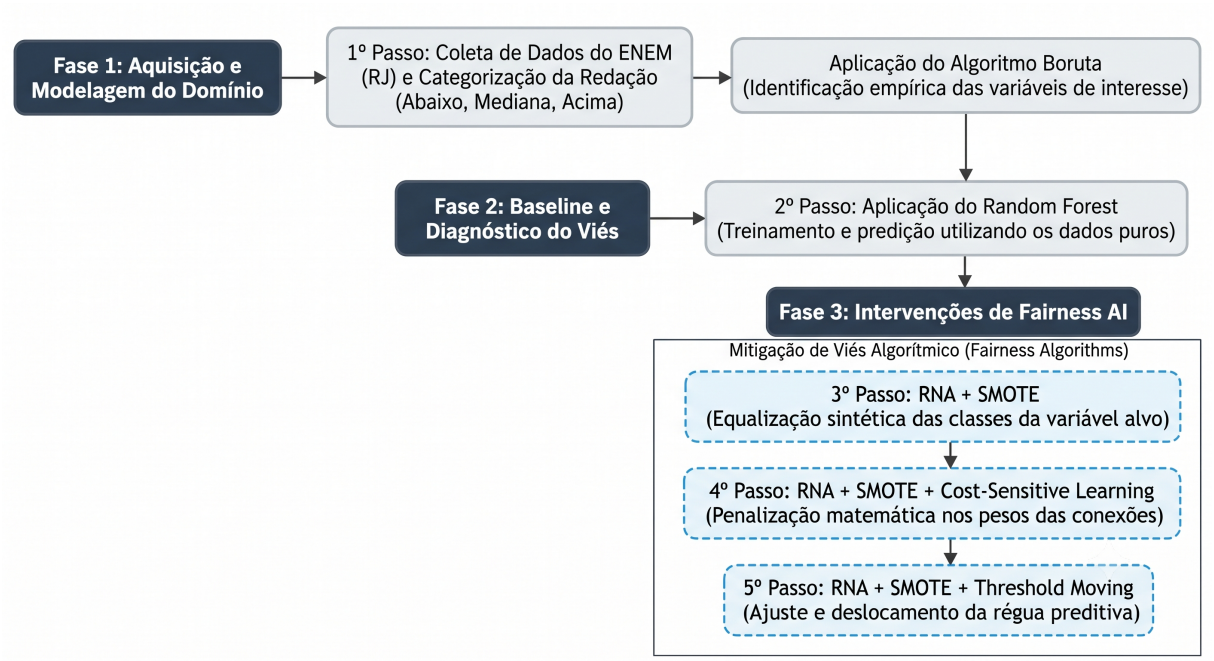
Este estudo caracteriza-se como uma investigação transversal de natureza descritivo-analítica, fundamentada em dados quantitativos secundários. No que tange aos seus objetivos, a pesquisa classifica-se, primariamente, como exploratória e explicativa. Quanto aos procedimentos técnicos adotados, o delineamento metodológico baseia-se em duas frentes distintas: a pesquisa documental e a pesquisa experimental (Gil, 2002; Medeiros et al., 2010).

Conforme ilustrado na Figura 1, o percurso metodológico desta pesquisa foi estruturado em três fases progressivas de análise e intervenção. A primeira fase, dedicada à Aquisição e Modelagem do Domínio, consistiu na coleta dos microdados do ENEM referentes ao estado do Rio de Janeiro, seguida pela categorização da variável alvo (a nota da redação) em três estratos de desempenho: "Abaixo", "Mediana" e "Acima". Na mesma etapa, aplicou-se o algoritmo Boruta para a identificação empírica e seleção das variáveis sociodemográficas de maior interesse (Kursa; Rudnicki, 2010). A partir desse refinamento, a segunda fase estabeleceu o *baseline* preditivo ao aplicar o algoritmo *Random Forest* sobre os dados puros(Hastie et al., 2009). Esse segundo passo serviu como modelo de referência para diagnosticar o viés estrutural original e mensurar o comportamento algorítmico antes de qualquer equalização corretiva.

A terceira e última fase do fluxograma concentrou-se na experimentação de intervenções de Justiça Algorítmica (*Fairness AI*), desenhadas para mitigar o viés observado (Barocas et al., 2019). Este escopo metodológico foi desdobrado em três passos aplicados a uma Rede Neural Artificial (RNA): o terceiro passo introduziu a técnica SMOTE-NC para a equalização sintética das classes da variável alvo (Fernández et al., 2018; Blagus; Lusa, 2013); o quarto passo adicionou o *Cost-Sensitive Lear-*

[5]. A base de dados está disponível em: `https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/microdados/enem`, e foi consultada na data de oito de julho de 2026.

Figura 1: Passos Metodológicos Adotados no Trabalho



Fonte: Autoria Própria.

ning, aplicando penalizações matemáticas diretas nos pesos das conexões da rede durante a fase de treinamento (Elkan, 2001); e, por fim, o quinto passo implementou o *Threshold Moving*, que realiza o ajuste e o deslocamento probabilístico preditivo no momento pós-treinamento(Zhou; Liu, 2006). Essa arquitetura metodológica permite avaliar de forma sistemática se as sucessivas correções matemáticas são capazes de solucionar a falha preditiva sobre os candidatos situados na zona mediana.

Em estrito cumprimento às diretrizes e boas práticas de Ciência Aberta, garantindo a total auditabilidade e reprodutibilidade dos achados empíricos aqui discutidos, todos os artefatos computacionais desta pesquisa foram unificados e depositados publicamente no repositório científico Zenodo. O compêndio metodológico engloba os recortes sociodemográficos anonimizados dos microdados do ENEM (Estado do Rio de Janeiro), os dicionários de variáveis do INEP e os *scripts* originais desenvolvidos em linguagem *Python*. O repositório documenta de ponta a ponta a higienização da base, o treinamento dos modelos preditivos (*Random Forest* e Rede Neural Artificial) e a aplicação das técnicas de correção de viés (*Cost-Sensitive Learning* e *Threshold Moving*). O conjunto completo de dados, códigos e resultados visuais está preservado de forma imutável e pode ser acessado integralmente, através de Freitas Neto et al. (2026)

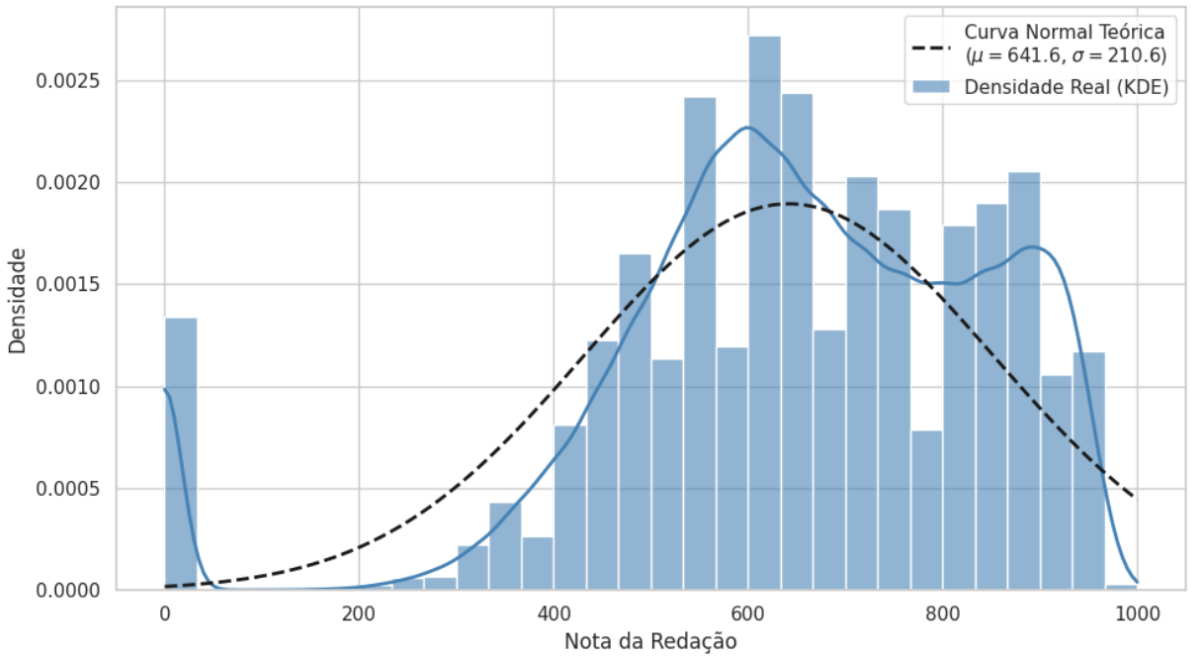
4 Resultados e Discussão

4.1 Perfil Estatístico Descritivo e a Distribuição da Variável Alvo

O tratamento do *dataset* desta pesquisa iniciou-se com a extração e análise descritiva da variável alvo (`'NU_NOTA_REDACAO'`), que corresponde à nota final da redação de 213.345 candidatos do Estado do Rio de Janeiro. O diagnóstico dos valores centrais da amostra revelou uma média de 641,59 pontos, acompanhada por um elevado desvio padrão de 210,63. A mediana estabeleceu-se em 640,00 pontos, enquanto a moda da população fluminense concentrou-se na marca de 600,00 pontos. A Figura 2 demonstra que, embora a proximidade matemática entre a média e a mediana possa sugerir uma certa estabilidade no letramento do estado, a observação gráfica do histograma revela a realidade material

das diferenças educacionais.

Figura 2: Distribuição das Notas de Redação (ENEM/RJ) com Ajuste da Curva Normal



Autor: Autoria Própria

Conforme apresentado na Figura 2, a sobreposição da densidade real (KDE) com a projeção da curva normal teórica evidencia assimetrias que refutam a premissa de distribuição igualitária do capital cultural. Destaca-se, de forma imediata, a anomalia estrutural na cauda inferior do gráfico: um pico expressivo e antinatural de notas zero. Esse agrupamento representa o traço demográfico de candidatos excluídos do processo de enunciação avaliativa, seja por desclassificação, fuga ao tema ou ausência sistêmica.

Adicionalmente, a linha de densidade afasta-se da curva teórica na região central, acumulando uma frequência alta de sujeitos em torno da moda (600 pontos) e desenhando platôs irregulares nas faixas de excelência. Do ponto de vista discursivo e sociológico, essa ausência de normalidade não constitui um ”ruído” a ser corrigido por transformações matemáticas para forçar a simetria do banco de dados, mas sim a quantificação rigorosa do abandono histórico. O distanciamento entre a curva teórica e a realidade fluminense é a própria métrica da desigualdade que o *Machine Learning* será desafiado a interpretar, validando a decisão metodológica de não tratar as notas como uma regressão contínua.

Para operacionalizar as intervenções algorítmicas, optou-se por converter esse espectro contínuo[6] de assimetrias em uma variável alvo categórica, delimitando matematicamente 3 zonas da população. O corte por tercis força o algoritmo a lidar com classes de volume demográfico perfeitamente idênticas (33,3% dos candidatos em cada fatia), estabelecendo a classe de desempenho ”Abaixo” para notas até 580, a classe ”Acima” para notas superiores a 740, e restringindo a classe ”Mediana” entre 580 e 740 pontos.

A adoção dos Tercis cumpre uma dupla função estratégica: computacionalmente, fornece ao modelo de *baseline* (*Random Forest*) um cenário de classes perfeitamente balanceadas em volume de dados, isolando o viés algorítmico do desbalanceamento amostral; sociologicamente, circunscreve a classe ”Mediana” à exata zona de transição do capital cultural, onde as variáveis de exclusão absoluta

[6]Na realidade, a Nota de Redação é um conjunto de categorias com intervalos de 20 pontos, devido à média de dois processos de correção de cinco categorias, cada uma representando o valor da nota como múltiplos de 40 entre zero e 200 pontos (Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira, 2024).

e de privilégio estrutural perdem seu peso determinístico e a complexidade do sujeito passa a não ser detectada pela máquina.

4.2 Seleção das Variáveis: A Aplicação do Algoritmo Boruta

Definida a variável alvo e a sua categorização matemática em tercis, a etapa seguinte exigiu a delimitação do espaço de entrada (*features*) que alimentaria os modelos preditivos. O questionário socioeconômico e as bases de inscrição do INEP fornecem dezenas de metadados sobre cada candidato, variando desde a cor/raça e renda familiar até a escolaridade dos pais e a posse de bens materiais. Optar por selecionar essas variáveis de forma manual ou arbitrária inseriria o viés do pesquisador na modelagem, além de ser impraticável devido a complexidade que estas intersecções provocam na observação do sujeito em ecossistema sociodemográfico.

Para garantir o rigor metodológico e a isenção na modelagem do domínio sociológico, aplicou-se o algoritmo Boruta (Kursa; Rudnicki, 2010). Diferentemente de métodos tradicionais de redução de dimensionalidade, o Boruta atua como um sistema empírico de validação. O algoritmo opera criando cópias exatas de todas as variáveis originais da base do Rio de Janeiro, denominadas *shadow features* (variáveis sombra), e embaralha os seus valores para testar a relação de causalidade com a nota da redação.

Em seguida, um modelo de *Random Forest* é treinado para extrair a métrica de importância (*Feature Importance*) de cada atributo, instaurando uma disputa: a variável real é comparada contra a melhor variável sombra. Apenas os marcadores sociodemográficos que superam sistematicamente o ruído aleatório em múltiplas iterações são confirmados e retidos. Essa filtragem algorítmica garante que os modelos preditivos das fases seguintes sejam alimentados exclusivamente com dados que possuem impacto comprovado sobre o desempenho do candidato, expurgando ruídos estatísticos. Dessa forma, a exclusão delineada pela máquina passa a ser pautada estritamente nas variáveis que a própria estatística reconheceu como os verdadeiros limitadores sociais do estado fluminense.

A execução do algoritmo resultou na retenção de 32 variáveis como preditores de alta importância, enquanto sete atributos foram estatisticamente rejeitados pelo modelo. A análise das colunas descartadas revela a precisão do Boruta em ignorar dados sem variância ou poder discriminatório. Variáveis como ano do exame (`NU_ANO_x`, `NU_ANO_y`) e identificadores estaduais (`CO_UF_PROVA`, `SG_UF_PROVA`) foram corretamente ignoradas, uma vez que a amostragem já estava restrita à edição de 2024 e ao território fluminense. Da mesma forma, a nacionalidade (`TP_NACIONALIDADE`) e questões muito específicas do questionário socioeconômico relativas a bens de consumo básicos (como `Q012` e `Q014`, referentes à posse de geladeira e máquina de lavar) foram rejeitadas, demonstrando que a posse de eletrodomésticos universais não configura um diferencial de letramento.

Em contrapartida, as 32 colunas escolhidas pelo algoritmo como preditores essenciais fornecem uma visão da classificação social. A máquina identificou que, para prever a nota de um candidato, ela precisa decodificar o seu capital econômico e cultural. Foram confirmadas variáveis estruturais como a escolaridade do pai e da mãe (`Q001` e `Q002`), a renda familiar mensal (`Q006`), o tipo de escola frequentada (`TP_ENSINO`) e marcadores identitários cruzados, como o sexo e a cor/raça do aluno (`TP_SEXO` e `TP_COR_RACA`). A inclusão do município de realização da prova (`CO_MUNICIPIO_PROVA` e `NO_MUNICIPIO_PROVA`) atesta, ainda, que o algoritmo percebeu a assimetria geográfica fluminense.

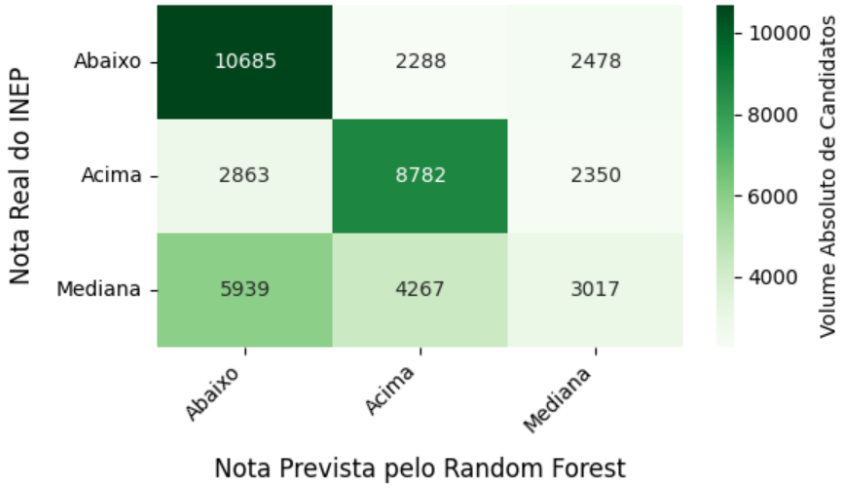
4.3 O Teste *Baseline*: Eficácia nos Extremos e o Incerteza na Zona Mediana

Com o espaço de entrada (*features*) rigorosamente filtrado pelo algoritmo Boruta, a segunda fase metodológica consistiu no treinamento do modelo de referência (*baseline*) utilizando o algoritmo

Random Forest sobre os dados puros. O objetivo desta etapa não era alcançar a precisão preditiva máxima, mas sim estabelecer um diagnóstico cru e algorítmico do viés estrutural fluminense antes de qualquer intervenção de mitigação ética.

O treinamento do algoritmo retornou uma acurácia global de 52,69%. Prever corretamente pouco mais da metade da base de dados poderia ser interpretado como um desempenho medíocre. No entanto, através da Matriz de Confusão (conforme ilustrado na Figura 3), confirma a tese central desta investigação: a acurácia do modelo não se distribui de maneira uniforme, mas gera incerteza ao encontrar a complexidade do sujeito mediano.

Figura 3: Matriz de Confusão: *Random Forest* vs. Base de Testes



Fonte Autoria Própria.

A matriz (Figura 3) demonstra que a Inteligência Artificial é eficaz na leitura dos extremos sociais. O algoritmo demonstrou acurácia ao classificar os candidatos da classe ”Abaixo”, acertando a predição de 10.685 indivíduos, assim como na classe ”Acima”, com 8.782 predições corretas. Esse fenômeno comprova que, nos extremos da distribuição sociodemográfica, o determinismo é capaz de representar a classificação. Quando a vulnerabilidade material é determinante, ou quando o privilégio estrutural é consolidado, os marcadores identitários e econômicos selecionados pelo Boruta atuam como vetores representativos. A máquina, nesses casos, não precisa ler a redação; ela deduz a nota por meio do peso histórico da exclusão ou do capital cultural acumulado.

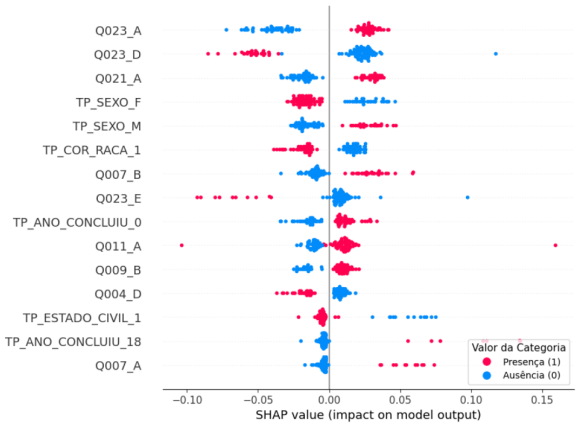
Em contrapartida, a linha correspondente à nota real ”Mediana” evidencia a dificuldade do modelo. Dos 13.223 candidatos que, de fato, pertenciam à zona central de desempenho, o *Random Forest* conseguiu identificar apenas 3.017. A maioria dos candidatos medianos foi dividida e classificada pelo algoritmo entre os extremos: 5.939 foram subestimados e alocados na classe ”Abaixo”, enquanto 4.267 foram superestimados e alocados na classe ”Acima”.

Para compreender os mecanismos internos que conduziram a essa indeterminação algorítmica, foram extraídos os valores SHAP (*SHapley Additive exPlanations*), os quais decodificam o peso exato de cada variável na decisão da máquina para cada estrato (conforme os gráficos apresentados na Figura 4) (Lundberg; Lee, 2017). A extração das contribuições marginais revelou como o algoritmo se apoia no determinismo estrutural para julgar os polos da sociedade, enquanto falha ao tentar parametrizar a complexidade e a subjetividade autoral da faixa mediana.

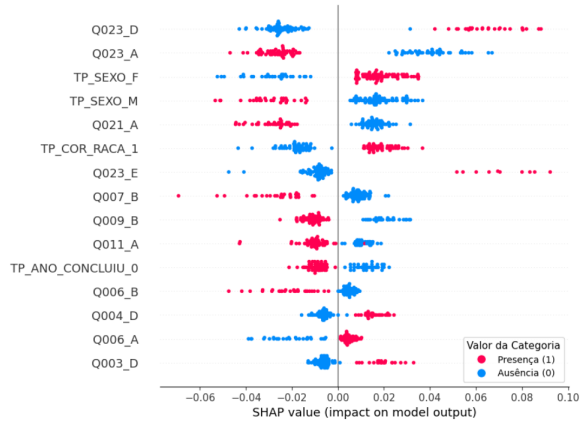
A análise visual dos gráficos *Beeswarm* correspondentes aos estratos extremos evidencia um espelhamento algorítmico perfeito. No espectro da vulnerabilidade ("Abaixo da Média", conforme Figura 4a), a presença de variáveis indicativas de vulnerabilidade, como o acesso exclusivo ao ensino público (`Q023_A`), projeta-se acentuadamente para a direita do eixo zero, assumindo um forte vetor de

exclusão que diminui a nota do candidato. Em contrapartida, no estrato de excelência ("Acima da Média", conforme Figura 4b), observa-se que a presença desses mesmos limitadores materiais exerce uma força negativa (projetando-se para a esquerda do eixo), afastando agressivamente o candidato da oportunidade de obter uma nota alta. Marcadores de privilégio estrutural e de cor/raça assumem um controle preditivo com vetores claros de aprovação.

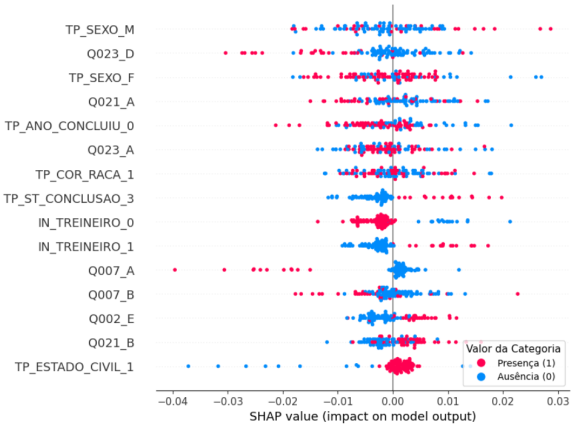
Figura 4: Distribuição dos valores SHAP demonstrando o peso de cada variável sociodemográfica na decisão preditiva do algoritmo para cada classe de letramento.



(a) Impacto no Estrato Abaixo.



(b) Impacto no Estrato Acima.



(c) Impacto no Estrato Mediana.

Fonte: Autoria Própria.

No entanto, o caso mais interessante ocorre no gráfico referente ao estrato central ("Classe Mediana", conforme Figura 4c). Ao contrário das classes extremas, onde os vetores são segregados e direcionais, a distribuição SHAP da zona mediana apresenta uma indeterminação. Observa-se que a presença ou ausência das características demográficas se mistura indistintamente em ambos os lados do eixo zero para quase todas as variáveis. Além disso, a escala do eixo X (Impacto SHAP) sofre uma drástica compressão: enquanto o peso das variáveis nas extremidades variava livremente em faixas amplas (entre -0.06 e 0.10), na classe mediana, nenhuma variável consegue exercer uma força superior à margem de $\pm$ 0.03.

4.4 Balanceamento Sintético (SMOTE-NC) e Redes Neurais (MLP) na Zona de Indeterminação

Diante da dificuldade do modelo de referência (*Random Forest*) ao tentar parametrizar a zona mediana, a terceira fase metodológica desta pesquisa recorreu a uma arquitetura de maior complexidade matemática: uma Rede Neural Artificial do tipo *Multilayer Perceptron* (MLP) (Haykin, 1999; Russell; Norvig, 2020), acoplada a uma intervenção ética de balanceamento por meio do SMOTE-NC (*Synthetic Minority Over-sampling Technique for Nominal and Continuous*) (Fernández et al., 2018; He; Garcia, 2009; Chawla et al., 2002). O objetivo estrutural desta etapa foi verificar se a capacidade de uma rede profunda em mapear relações não-lineares, somada à equalização estatística das classes, conseguiria decodificar a complexidade envolvida neste estrato intermediário de notas.

Contudo, a manipulação ética e a inserção de dados artificiais exigem protocolos estritos para não comprometer a validade epistemológica da pesquisa. A literatura de Ciência de Dados adverte que a geração de registros sintéticos, se mal particionada, pode causar o vazamento de informações (*data leakage*) (Vandewiele et al., 2021). Se os dados gerados artificialmente pelo SMOTE-NC contaminarem o ambiente de teste, a rede neural será avaliada com base em exemplos que ela mesma ajudou a criar, resultando em uma acurácia ilusória e superestimada.

Para impedir a contaminação da avaliação e evitar o vazamento de dados, foi estabelecida uma separação da amostra. A população de 213.345 registros foi dividida em três partições isoladas e mutuamente exclusivas, garantindo que o aprendizado, a validação e o teste final ocorressem em um conjunto de dados separados, conforme detalhado no Quadro 2.

Quadro 2 – Particionamento do *corpus* e arquitetura de isolamento metodológico.

Partição	Volume	Uso do SMOTE-NC	Função Metodológica
Treinamento	136.540 (64%)	**Sim**	Único estrato com equalização sintética para nivelar perfeitamente o volume das três classes (‘Abaixo’, ‘Acima’ e ‘Mediana’), garantindo que a rede neural aprenda os padrões de todas as faixas de letramento sem o viés estatístico das maiorias.
Validação (RNA)	34.136 (16%)	**Não**	Fração de controle interno. Utilizada a cada época (*epoch*) do treinamento para balizar o ajuste de pesos da rede e prevenir o sobreajuste (*overfitting*).
Teste Final	42.669 (20%)	**Não**	Atua como prova definitiva para aferir a capacidade de generalização do modelo no “mundo real”.

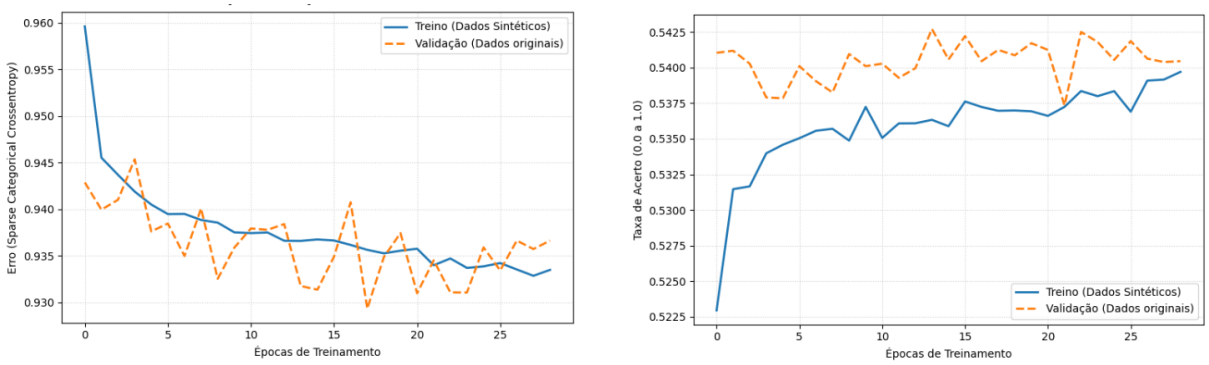
Fonte: Autoria Própria.

A topologia do *Multilayer Perceptron* (MLP) foi estabelecida após o processo de otimização de hiperparâmetros, conduzido por meio da biblioteca *Keras Tuner* (O’Malley et al., 2019). A arquitetura computacional foi instanciada utilizando o modelo Sequencial, resultando na seguinte configuração técnica estrutural:

- Primeira Camada Oculta: Instanciada com 32 neurônios, essa camada recebe a dimensionalidade dos dados de entrada (`input_dim`) e é configurada para utilizar a função de ativação ReLU (*Rectified Linear Unit*) (Haykin, 1999).
- Camada de Regularização: Implementa uma operação de *Dropout* com taxa de 30% (0.3) imediatamente após a primeira camada, com o objetivo de prevenir o sobreajuste (*overfitting*) (Silva et al., 2010; Haykin, 1999).
- Segunda Camada Oculta: Composta por 32 neurônios e configurada com a função de ativação ReLU, esta camada atua na consolidação dos padrões processados.
- Camada de Saída: Estruturada com 3 neurônios — correspondentes às classes preditivas do experimento (”Abaixo”, ”Acima” e ”Mediana”) — e submetida à função de ativação *Softmax*, responsável por converter o processamento em um vetor de probabilidades cuja soma totaliza 100% (Silva et al., 2010).

Para monitorar a convergência do MLP e atestar a eficácia, o comportamento da rede foi rastreado ao longo das épocas de treinamento (*epochs*). A Figura 5 ilustra o histórico de otimização do modelo, contrastando o desempenho da rede em um ambiente controlado pelo SMOTE-NC com a generalização nos dados originais de validação.

Figura 5: Curvas de aprendizado da Rede Neural contrastando os dados sintéticos (treino) com a realidade assintótica (validação).



(a) Evolução da Função de Perda (Loss).

(b) Evolução da Acurácia Global.

Fonte: Autoria Própria.

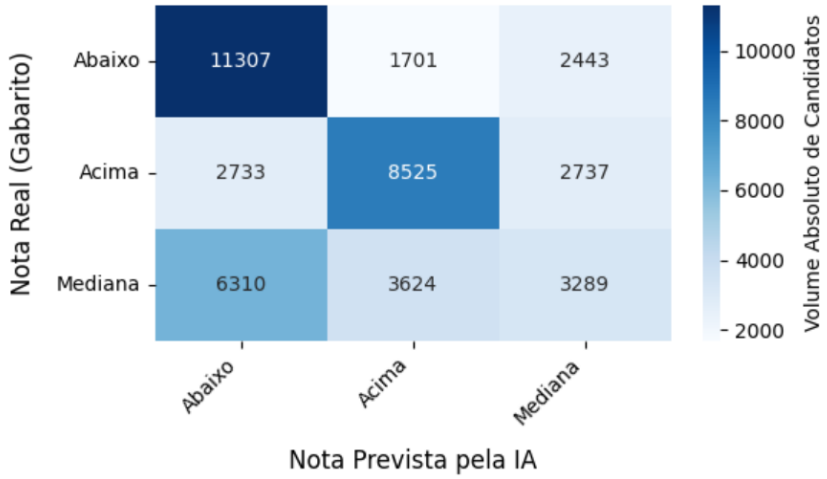
A análise da Curva de Erro (*Sparse Categorical Crossentropy*), apresentada na Figura 5a, evidencia uma queda nas primeiras épocas de treinamento, seguida de estabilização na faixa de 0,935. A linha de validação, embora apresente um comportamento oscilatório natural, decorrente do confronto entre a rede (treinada em dados sintéticos e equalizados) e os dados reais, acompanha a tendência de descida do treino sem divergir para cima. Esse comportamento demonstra o sucesso técnico da barreira ética *Dropout* (30%): a rede não memorizou os dados sintéticos, evitando o fenômeno do sobreajuste (*overfitting*).

O gráfico de Acurácia Global, (vide Figura 5b, que apresenta o valor inicial do aprendizado em torno de 52% e, à medida que ajusta seus pesos sinápticos, estabiliza-se em torno de 54%. A acurácia de validação evolui nessa mesma margem.

Essa paralisação preditiva representa um limite algorítmico. O *Baseline* inicial, utilizando o algoritmo *Random Forest*, havia alcançado 52,69%. Ao substituí-lo por uma arquitetura profunda (MLP), capaz de decodificar relações interseccionais e não-lineares, e ao neutralizar o desbalanceamento numérico com o SMOTE-NC, obteve-se um ganho de performance global que foi inferior a dois pontos percentuais.

A Figura 6 apresenta as predições do MLP, treinado sob o nivelamento sintético do SMOTE-NC, revelando de que forma a máquina classifica os candidatos ao se deparar com a complexidade dos dados intersecionais do ENEM.

Figura 6: Matriz de Confusão: MLP vs. Base de Teste



Fonte: Autoria Própria.

A leitura das linhas da matriz (Nota Real), em contraste com as colunas (Nota Prevista pela IA), demonstra que o algoritmo manteve sua acurácia nos extremos estruturais. Dos 15.451 candidatos que de fato pertenciam ao estrato ”Abaixo”, o MLP acertou a classificação de 11.307 (aproximadamente 73,1% de acurácia isolada nesta classe). No estrato ”Acima”, dos 13.995 candidatos do estrato ”Acima”, o modelo classificou corretamente 8.525 indivíduos (60,9%). Reitera-se que, nas faixas de exclusão ou de excelência garantida pelas condições materiais, os dados sociodemográficos operam como uma sentença estatística previsível pela máquina.

Entretanto, no estrato correspondente aos 13.223 candidatos reais da classe ”Mediana”. Apesar de o SMOTE-NC ter treinado a rede neural com um volume massivo e equilibrado de exemplos, a máquina enfrentou uma incapacidade preditiva ao tentar identificá-los no mundo real. Dos mais de 13 mil candidatos medianos, o MLP acertou apenas 3.289 (uma taxa de acerto de 24,8%), inferior até mesmo a probabilidade aleatória de 33% em um cenário de três classes. A classificação atribuída pelo algoritmo a esses candidatos evidencia seu viés estrutural: a rede comprimiu a classe central, dispersando seus indivíduos para os polos. Um total de 6.310 candidatos medianos foram classificados na categoria ”Abaixo”, enquanto 3.624 foram alocados na categoria ”Acima”.

4.5 A Aplicação do *Cost-Sensitive Learning* na Zona de Indeterminação

Os resultados observados na classificação do estrato mediano indicaram a necessidade de revisar as estratégias de mitigação do viés algorítmico. A aplicação prévia do SMOTE-NC sugeriu que a dificuldade do modelo em identificar os padrões da classe central não decorre estritamente de uma sub-representação quantitativa na base de dados. Observou-se que o balanceamento sintético, por si só, não foi suficiente para elevar, de forma significativa, a acurácia nessa faixa específica. Diante disso, a abordagem metodológica para a busca por Justiça Algorítmica (*Fairness AI*) foi redirecionada da manipulação do *corpus* de entrada para a adequação da função de otimização da Rede Neural.

Para atenuar a tendência do modelo de dispersar os candidatos medianos para os estratos extremos, adotou-se a técnica de *Cost-Sensitive Learning* (Aprendizado Sensível ao Custo). Em arquiteturas padrão de aprendizado de máquina, a função de perda (*Loss Function*) geralmente atribui pesos

equivalentes a todas as classificações incorretas, independentemente da classe analisada. A abordagem sensível ao custo modifica essa dinâmica ao incorporar uma matriz de penalidades assimétricas durante a fase de treinamento da máquina (Elkan, 2001; Kamiran; Calders, 2012).

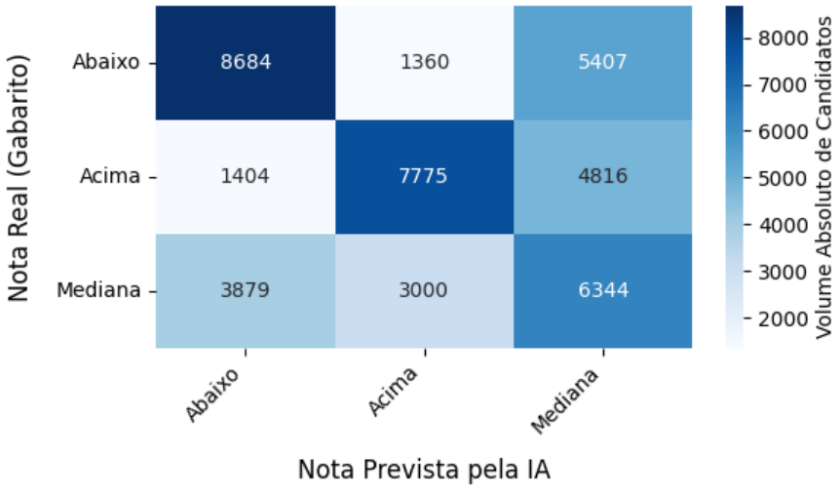
Com essa nova configuração, o algoritmo passa a considerar o impacto desproporcional dos erros preditivos. Ao identificar que a rede apresenta uma propensão a gerar falsos positivos nos polos demográficos em detrimento do estrato mediano, o modelo é ajustado para aplicar um peso maior ao erro (um multiplicador na função de perda) sempre que ocorrer a classificação incorreta desse grupo específico. O objetivo dessa intervenção é induzir o algoritmo a priorizar a extração dos padrões e das interseccionalidades da classe central, adequando a tomada de decisão sem a necessidade de introduzir novas instâncias artificiais nos dados originais.

Para configurar essa intervenção no processo de aprendizado, a matriz de pesos preditivos (*class_weight*) foi parametrizada de forma a refletir a dificuldade algorítmica observada nos testes anteriores. Aos estratos demográficos extremos: ”Abaixo” (Classe 0) e ”Acima” (Classe 1) — atribuiu-se o peso unitário e neutro de 1.0, estabelecendo a linha de base para o cálculo da função de perda. Em contrapartida, à classe ”Mediana” (Classe 2) foi atribuído um peso de 1,3.

Matematicamente, essa configuração estabelece que qualquer erro classificatório envolvendo um candidato da zona central resulte em um incremento de 30% no valor do erro (*loss*) processado pela rede. O mecanismo de retropropagação (*backpropagation*) é, portanto, obrigado a ajustar as conexões sinápticas com maior ênfase nos traços dessa população, a fim de minimizar essa penalidade acumulada.

A avaliação do modelo refinado pode ser vista na Figura 7, que detalha o comportamento preditivo da Rede Neural Artificial após a injeção de um peso adicional de 30% na função de perda da classe ”Mediana”.

Figura 7: Matriz de Confusão: MLP - *Cost-Sensitive-Learning* vs. Base de Teste



Fonte: Autoria Própria.

A análise comparativa da nova Matriz de Confusão indica a ocorrência de um fenômeno de compensação algorítmica (*trade-off*), no qual o aumento da sensibilidade (*recall*) para a classe mal representada resulta na degradação acelerada da precisão global do sistema. Ao observar a taxa de acertos da classe ”Mediana” (linha 3), constata-se um aumento quantitativo expressivo: o modelo identificou corretamente 6.344 indivíduos, quase o dobro do que foi observado na arquitetura não penalizada.

Contudo, esse avanço local foi obtido à custa da desestabilização dos polos preditivos. A matriz evidencia que o algoritmo, ao tentar minimizar a penalidade matemática configurada na matriz de cus-

tos, passou a direcionar instâncias fronteiriças para a classe ”Mediana”. Esse desvio é observável no volume de falsos positivos alocados na terceira coluna: a rede classificou incorretamente 5.407 sujeitos do estrato "Abaixo"e 4.816 do estrato "Acima"como pertencentes à classe mediana.

Consequentemente, o acerto determinístico que o modelo apresentava em relação aos extremos estruturais foi comprometido. A identificação correta da classe ”Abaixo” recuou de 11.307 para 8.684, enquanto o reconhecimento do estrato ”Acima” caiu de 8.525 para 7.775. A acurácia global do sistema sofreu um decréscimo, comprovando que a tentativa de forçar a convergência estatística, neste caso, não resolveu a previsibilidade da informação.

4.6 O Deslocamento de Limiares Preditivos (*Threshold Moving*) na Zona de Indeterminação

O aprendizado sensível a custos (*Cost-Sensitive Learning* - CSL), quando aplicado a este problema, evidenciou que a intervenção direta na função de perda durante a fase de treinamento desestabiliza o reconhecimento dos polos sociodemográficos. Ao tentar forçar a rede neural a extrair padrões inexistentes no vetor demográfico, o modelo comprometeu sua precisão nas extremidades. Por conseguinte, a última fase metodológica deste experimento abdica da alteração dos pesos sinápticos (CSL) em favor de uma técnica de pós-processamento: o *Threshold Moving* (Deslocamento de Limiar) (Sheng; Ling, 2006).

Diferentemente das abordagens de penalização interna, o *Threshold Moving* opera sob a premissa de que a arquitetura neural, alimentada pela base balanceada, extraiu a máxima informação possível do *corpus*. Sob essa ótica, torna-se metodologicamente mais seguro manter a topologia da rede intacta e intervir apenas na interpretação matemática de sua saída (Sheng; Ling, 2006). No comportamento algorítmico padrão, a classificação final de um candidato é determinada pela regra do argumento máximo (*argmax*): a classe que obtém a maior probabilidade bruta no vetor gerado pela função *Softmax* é eleita como o veredito do modelo. O deslocamento de limiar atua recalibrando essa regra de decisão de forma localizada.

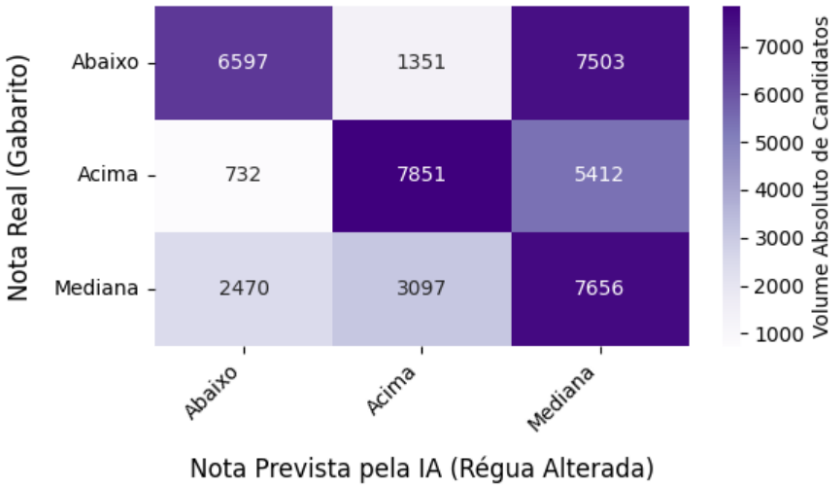
Ao aplicar essa técnica no contexto desta pesquisa, reduz-se a exigência probabilística mínima para que o algoritmo classifique uma instância no estrato ”Mediana”. O limiar de aceitação para esta classe foi rebaixado para 34% (0,34)[7]. Essa intervenção reconhece formalmente as limitações do modelo que opera na zona de indeterminação. Em vez de exigir que a máquina alcance uma elevada certeza estatística fundamentada em *proxies* demográficas ambíguas, o sistema é ajustado para aceitar sinais probabilísticos mais sutis.

A observação da matriz ajustada (vide Figura 8) revela que a intervenção atingiu seu objetivo primário: a identificação correta dos candidatos do estrato mediano obteve o melhor desempenho absoluto em todo o experimento, alcançando 7.656 instâncias de verdadeiras positivas. Contudo, a flexibilização da fronteira de decisão evidenciou a sobreposição estatística que caracteriza os tensores demográficos dessa população.

Ao reduzir o nível de exigência para a classificação na faixa central, o modelo absorveu um volume massivo de instâncias vizinhas. A terceira coluna da matriz (predição ”Mediana”) tornou-se um repositório que agrupou 7.503 falsos positivos originários do estrato ”Abaixo” e 5.412 candidatos do estrato ”Acima”. Consequentemente, a precisão nas extremidades sociodemográficas sofreu um recuo:

[7] Ressalta-se que a determinação do novo limiar de 34% (0,34) não ocorreu de forma arbitrária, mas resultou de um ajuste empírico e iterativo de hiperparametrização focado em maximizar o *recall* da classe central. Para resguardar o rigor metodológico e evitar o vazamento de dados, essa calibragem do ponto de corte foi executada e validada exclusivamente sobre a partição de Validação (16% da amostra). Uma vez estabelecido o ponto ótimo empírico de compromisso, o limiar foi congelado e aplicado de forma cega à partição de Teste Final (20%), garantindo que a avaliação da estagnação algorítmica refletisse a capacidade real de generalização do modelo diante de dados inéditos.

Figura 8: Matriz Ajustada (Threshold: 34%) vs. Base de Teste



Fonte: Autoria Própria.

a rede classificou corretamente 6.597 candidatos da classe inferior e 7.851 da classe superior.

5 Conclusão

Esta pesquisa propõe investigar os limites teóricos e operacionais da Inteligência Artificial na predição do desempenho discursivo humano no processo de avaliação da redação do ENEM de 2024, utilizando marcadores sociodemográficos como variáveis independentes. A hipótese central sustentada é que, embora algoritmos de *Machine Learning* sejam altamente eficientes em mapear o determinismo estrutural, eles encontram uma barreira de difícil reconhecimento computacional ao tentar decodificar a complexidade humana do letramento, onde a agência subjetiva e a polifonia se sobrepõem ao contexto material.

A trajetória metodológica adotada forneceu os recursos necessários para comprovar essa proposição. A evolução do experimento, partindo de um modelo *Baseline* (*Random Forest*), avançando para uma arquitetura profunda (*Multilayer Perceptron*) equalizada pelo algoritmo SMOTE-NC, e culminando em intervenções de Justiça Algorítmica (*Cost-Sensitive Learning* e *Threshold Moving*), demonstrou que a incapacidade da máquina em prever a nota em uma zona ”Mediana” não é uma falha de parametrização ou uma sub-representação estatística.

Entre as quatro abordagens testadas, nenhuma superou o teto de aproximadamente 54% de acurácia global, alcançado pela arquitetura MLP com balanceamento sintético (SMOTE-NC). As intervenções subsequentes de Justiça Algorítmica, embora tenham melhorado substancialmente o *recall* da classe ”Mediana” (o *Cost-Sensitive Learning* elevou os acertos nessa classe de 3.289 para 6.344, e o *Threshold Moving* os elevou ainda mais, para 7.656), foram obtidas à custa de um recuo em relação ao teto de 54,18% (53,44% e 51,79%, respectivamente), com o *Threshold Moving* regredindo inclusive abaixo do *baseline* original (52,69%). Esse padrão evidencia um *trade-off* sistemático entre justiça algorítmica localizada e desempenho agregado: ao flexibilizar as regras de decisão para acomodar o estrato mediano, observa-se não a superação do teto preditivo, mas a redistribuição matemática do erro entre as classes, com a categoria central passando a absorver falsos positivos oriundos das extremidades. Esse fenômeno comprova que os tensores demográficos de um candidato precarizado que ascende à nota mediana e de um candidato privilegiado que sofre um revés avaliativo tornam-se, na fronteira da zona de indeterminação, estatisticamente indistinguíveis. Além disso, nenhuma das técni-

cas de mitigação de viés aqui testadas é capaz de resolver essa indistinguibilidade sem comprometer a precisão nos extremos.

É nesse limite quantitativo que a Análise Preditiva deve ceder espaço a Análise Discursiva. O resíduo estatístico, ou seja, a margem de erro que o algoritmo não conseguiu mitigar, representa o espaço exato de operação exclusiva da autoria. Como os modelos adotados aqui não possuem compreensão semântica e operam exclusivamente por meio de correlações estocásticas de *proxies* estruturais (CEP, renda, escolaridade parental), eles permanecem, portanto, estruturalmente incapazes de identificar o esforço discursivo individual. O sujeito mediano utiliza seu repertório, sua voz e sua capacidade de articulação textual para subverter os padrões modelados de contextos socioeconômicos.

Portanto, conclui-se que, sob condições sociais igualitárias, o letramento e a autoria humana apresentam propriedades não-lineares que resistem à redução tensorial. Se os limites computacionais apresentados por Morin estabeleceram as fronteiras ontológicas do que pode ser modelado por uma máquina, os achados deste estudo demonstram que a subjetividade discursiva, teorizada por Bakhtin, habita fora desse escopo computável. O fracasso preditivo do algoritmo na classe ”Mediana” de notas é, em última análise, a prova empírica de que o sujeito enunciador não é apenas o produto de suas contingências estatísticas, mas também o autor de sua própria enunciação.

Declaração de disponibilidade de dados de pesquisa

Os conjuntos de dados estruturados e os códigos-fonte (scripts em Python) que dão suporte aos resultados empíricos deste estudo estão arquivados de forma imutável e disponibilizados publicamente no repositório científico Zenodo, podendo ser acessados sob o DOI: `https://doi.org/10.5281/zenodo.21937113`. Os microdados brutos originais utilizados como base primária são de domínio público e encontram-se disponíveis no Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira (2024).

Contribuição de autoria

- Francisco Alves de Freitas Neto: Conceituação, Curadoria de dados, Análise Formal, Investigação, Metodologia, Software, Visualização, Escrita – rascunho original, Escrita – revisão e edição;
- Carlos Henrique Medeiro de Souza: Conceituação, Supervisão, Validação, Escrita – revisão e edição;
- Fabio Machado de Oliveira: Metodologia, Validação, Escrita – revisão e edição;

Conflito de interesses

Os autores declaram não haver conflitos de interesses financeiros, comerciais, políticos, acadêmicos ou pessoais que possam ter influenciado a condução desta pesquisa, os resultados empíricos obtidos ou as interpretações apresentadas neste manuscrito.

Declaração de uso de Inteligência Artificial

Os autores declaram a utilização do modelo de inteligência artificial generativa Gemini (desenvolvido pelo Google) durante a elaboração deste trabalho. A ferramenta foi empregada com o escopo restrito de auxiliar na estruturação e formatação de linguagem de marcação (LaTeX), na revisão sintática e tradução de trechos específicos (como o *abstract*), e no refinamento e revisão das sintaxes

de programação em linguagem Python. Os autores atestam que revisaram e validaram rigorosamente todo o conteúdo gerado, assumindo total responsabilidade intelectual, técnica e ética pela autoria, bem como pelos dados, análises e conclusões apresentadas no manuscrito final.

Referências

ALYAHYAN, E.; DÜŞTEGÖR, D. Predicting academic success in higher education: literature review and best practices. **International Journal of Educational Technology in Higher Education**, SpringerOpen, v. 17, n. 1, p. 1–21, 2020.

AUTHIER-REVUZ, J. Heterogeneidade (s) enunciativa (s). **Cadernos de estudos linguísticos**, v. 19, p. 25–42, 1990.

BARDIN, L. **Análise de conteúdo**. Tradução: Luís Antero Reto e Augusto Pinheiro. São Paulo: Edições 70, 2011.

BAROCAS, S.; HARDT, M.; NARAYANAN, A. **Fairness and Machine Learning: Limitations and Opportunities**. [*S. l.*]: fairmlbook.org, 2019.

BARTHES, R. A morte do autor. *In:* O rumor da língua. Tradução: Mário Laranjeira. São Paulo: Martins Fontes, 2004. p. 57–64.

BLAGUS, R.; LUSA, L. SMOTE for high-dimensional class-imbalanced data. **BMC Bioinformatics**, BioMed Central, v. 14, n. 1, p. 1–15, 2013.

BODART, C. d. N.; COIMBRA, D. T. S. O Exame Nacional do Ensino Médio (ENEM) e a reprodução das desigualdades sociais. **Revista Brasileira de Estudos Pedagógicos**, SciELO Brasil, v. 101, p. 671–689, 2020.

CHAWLA, N. V.; BOWYER, K. W.; HALL, L. O.; KEGELMEYER, W. P. SMOTE: Synthetic Minority Over-sampling Technique. **Journal of Artificial Intelligence Research**, v. 16, p. 321–357, 2002.

ELKAN, C. The foundations of cost-sensitive learning. *In:* PROCEEDINGS of the 17th International Joint Conference on Artificial Intelligence (IJCAI). [*S. l.: s. n.*], 2001. v. 17, p. 973–978.

FEIJÓ, J. R.; FRANÇA, J. M. S. de. Diferencial de desempenho entre jovens das escolas públicas e privadas. **Revista Brasileira de Economia**, SciELO Brasil, v. 74, p. 471–490, 2020.

FERNÁNDEZ, A.; GARCÍA, S.; HERRERA, F.; CHAWLA, N. V. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary. **Journal of Artificial Intelligence Research**, v. 61, p. 863–905, 2018.

FIORIN, J. L. **Introdução ao pensamento de Bakhtin**. São Paulo: Contexto, 2016.

FOUCAULT, M. **O que é um autor?** 3. ed. Lisboa: Vega, 1997. (Passagens). ISBN 972-699-303-2.

FREITAS NETO, F. A. d.; SOUZA, C. H. M. d.; OLIVEIRA, F. M. d.; MOREIRA, L. B. **Dataset e Scripts para A Indeterminação da Autoria: A Modelagem Sociodemográfica e os Limites da Inteligência Artificial**. [*S. l.*]: Zenodo, 2026. Conjunto de dados e software. Acesso em: 14 ago. 2026. DOI: 10.5281/zenodo.21937113. Disponível em: https://doi.org/10.5281/zenodo.21937113.

FRIGOTTO, G.; CIAVATTA, M.; RAMOS, M. (ed.). **Ensino Médio integrado: concepção e contradições**. 3. ed. São Paulo: Cortez, 2012.

GIL, A. C. **Como elaborar projetos de pesquisa**. 4. ed. São Paulo: Atlas, 2002.

HASTIE, T.; TIBSHIRANI, R.; FRIEDMAN, J. **The Elements of Statistical Learning: Data Mining, Inference, and Prediction**. 2. ed. New York: Springer Science & Business Media, 2009.

HAYKIN, S. **Neural Networks: A Comprehensive Foundation**. 2nd. New York: Prentice Hall, 1999.

HE, H.; GARCIA, E. A. Learning from Imbalanced Data. **IEEE Transactions on Knowledge and Data Engineering**, v. 21, n. 9, p. 1263–1284, 2009.

INSTITUTO NACIONAL DE ESTUDOS E PESQUISAS EDUCACIONAIS ANÍSIO TEIXEIRA. **A redação no Enem 2024: cartilha do participante**. Brasília: Inep, 2024. Disponível em: https://download.inep.gov.br/publicacoes/institucionais/avaliacoes_e_exames_da_educacao_basica/a_redacao_no_enem_2024_cartilha_do_participante.pdf.

KAMIRAN, F.; CALDERS, T. Data preprocessing techniques for classification without discrimination. **Knowledge and Information Systems**, Springer, v. 33, n. 1, p. 1–25, 2012.

KURSA, M. B.; RUDNICKI, W. R. Feature selection with the Boruta package. **Journal of Statistical Software**, v. 36, n. 11, p. 1–13, 2010.

LUNDBERG, S. M.; LEE, S.-I. A unified approach to interpreting model predictions. *In:* ADVANCES in Neural Information Processing Systems. [*S. l.: s. n.*], 2017. v. 30.

MEDEIROS, C. H.; KAUARK, F. d. S.; MANHÃES, F. C. **Metodologia de Pesquisa: Um guia prático**. [*S. l.*]: Via Litterarum, Itabuna - BA, 2010. v. 1.

MORIN, E.; LE MOIGNE, J.-L. **A inteligência da complexidade**. São Paulo: Peirópolis, 2000. (Série nova consciência). ISBN 85-85663-42-1.

O'MALLEY, T.; BURSZTEIN, E.; LONG, J.; CHOLLET, F.; JIN, H.; INVERNIZZI, L. et al. **KerasTuner**. [*S. l.: s. n.*], 2019. GitHub repository. Disponível em: https://github.com/keras-team/keras-tuner.

O'NEIL, C. **Algoritmos de destruição em massa: como o big data aumenta a desigualdade e ameaça a democracia**. Santo André: Editora Rua do Sabão, 2020.

PIRES, A. Renda familiar e escolaridade dos pais: reflexões a partir dos microdados do Enem 2012 do Estado de São Paulo. **Revista de Ciências Humanas**, v. 51, n. 1, p. 152–173, 2017.

RASTROLLO-GUERRERO, J. L.; GÓMEZ-PULIDO, J. A.; DURÁN-DOMÍNGUEZ, A. Analyzing and predicting students' performance by means of machine learning: A review. **Applied Sciences**, MDPI, v. 10, n. 3, p. 1042, 2020. DOI: 10.3390/app10031042.

RUSSELL, S. J.; NORVIG, P. **Artificial Intelligence: A Modern Approach**. 4. ed. Hoboken, NJ: Pearson, 2020.

SHENG, V. S.; LING, C. X. Thresholding for making classifiers cost-sensitive. *In:* AAAI PRESS. PROCEEDINGS of the 21st National Conference on Artificial Intelligence - Volume 1. [*S. l.: s. n.*], 2006. p. 476–481.

SILVA, I. N. d.; SPATTI, D. H.; FLAUZINO, R. A. **Redes neurais artificiais para engenharia e ciências aplicadas: fundamentos teóricos e aspectos práticos**. 1. ed. São Paulo: Artliber, 2010.

VANDEWIELE, G.; DEHAENE, I.; KOVÁCS, G.; STERCKX, S.; JANSSENS, O.; ONGENAE, F.; DE BACKER, F.; DE TURCK, F.; ROELENS, K.; HOECKE, S. van. Overly optimistic prediction results on imbalanced data: a case study of flaws and benefits when applying over-sampling. **Artificial Intelligence in Medicine**, Elsevier, v. 111, p. 101987, 2021.

XAVIER, C. C.; RODRIGUES, C. K. d. S. Bridging the Connectivity Gap: A Multi-Year Study of Regional Inequities and Academic Outcomes in Brazil's ENEM (2015–2023). **Education Sciences**, MDPI, v. 14, n. 2, p. 1–22, 2024.

ZHOU, Z.-H.; LIU, X.-Y. Training cost-sensitive neural networks with methods addressing the class imbalance problem. **IEEE Transactions on Knowledge and Data Engineering**, IEEE, v. 18, n. 1, p. 63–77, 2006.

Este preprint foi submetido sob as seguintes condições:

- Os autores declaram que os necessários Termos de Consentimento Livre e Esclarecido de participantes ou pacientes na pesquisa foram obtidos e estão descritos no manuscrito, quando aplicável.
- Os autores declaram que a elaboração do manuscrito seguiu as normas éticas de comunicação científica.
- Os autores declaram que estão cientes que são os únicos responsáveis pelo conteúdo do preprint e que o depósito no SciELO Preprints não significa nenhum compromisso de parte do SciELO, exceto sua preservação e disseminação.
- Os autores declaram que os dados, aplicativos e outros conteúdos subjacentes ao manuscrito estão referenciados.
- O manuscrito depositado está no formato PDF.
- Os autores declaram que a pesquisa que deu origem ao manuscrito seguiu as boas práticas éticas e que as necessárias aprovações de comitês de ética de pesquisa, quando aplicável, estão descritas no manuscrito.
- Os autores declaram que uma vez que um manuscrito é postado no servidor SciELO Preprints, o mesmo só poderá ser retirado mediante pedido à Secretaria Editorial do SciELO Preprints, que afixará um aviso de retratação no seu lugar.
- Os autores concordam que o manuscrito aprovado será disponibilizado sob licença Creative Commons CC-BY.
- O autor submissor declara que as contribuições de todos os autores e declaração de conflito de interesses estão incluídas de maneira explícita e em seções específicas do manuscrito.
- Os autores declaram que o manuscrito não foi depositado e/ou disponibilizado previamente em outro servidor de preprints ou publicado em um periódico.
- Caso o manuscrito esteja em processo de avaliação ou sendo preparado para publicação mas ainda não publicado por um periódico, os autores declaram que receberam autorização do periódico para realizar este depósito.
- O autor submissor declara que todos os autores do manuscrito concordam com a submissão ao SciELO Preprints.